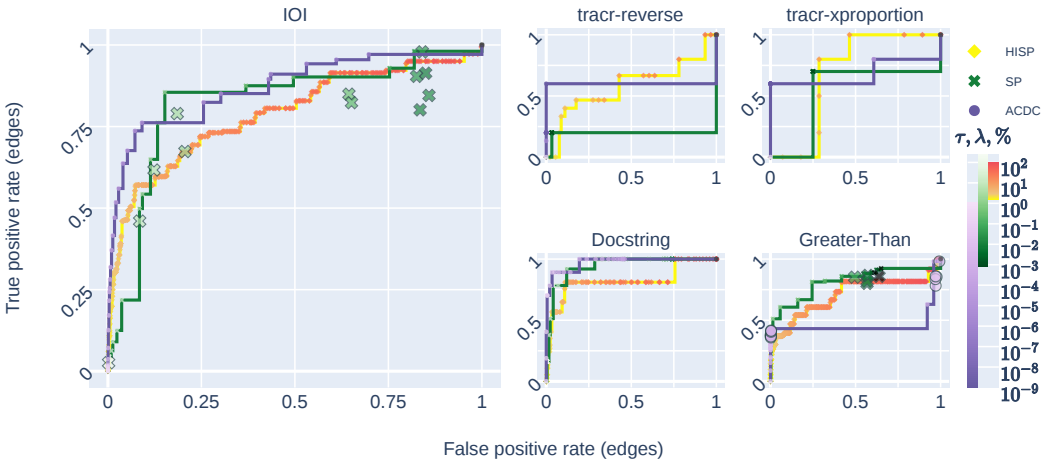


Figure 3: ROC curves of ACDC, SP and HISP identifying model components from previous work, across 5 circuits in transformers. We used target metrics from Table 1 for all tasks except ACDC, for which we used KL Divergence. The points on the plot are cases where SP and ACDC return subgraphs that are not on the Pareto Frontier. The full table of results is in Appendix J.

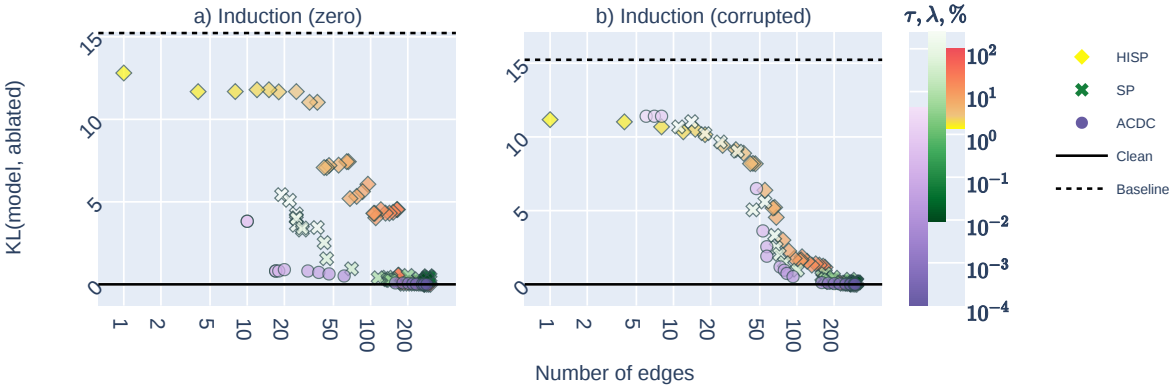
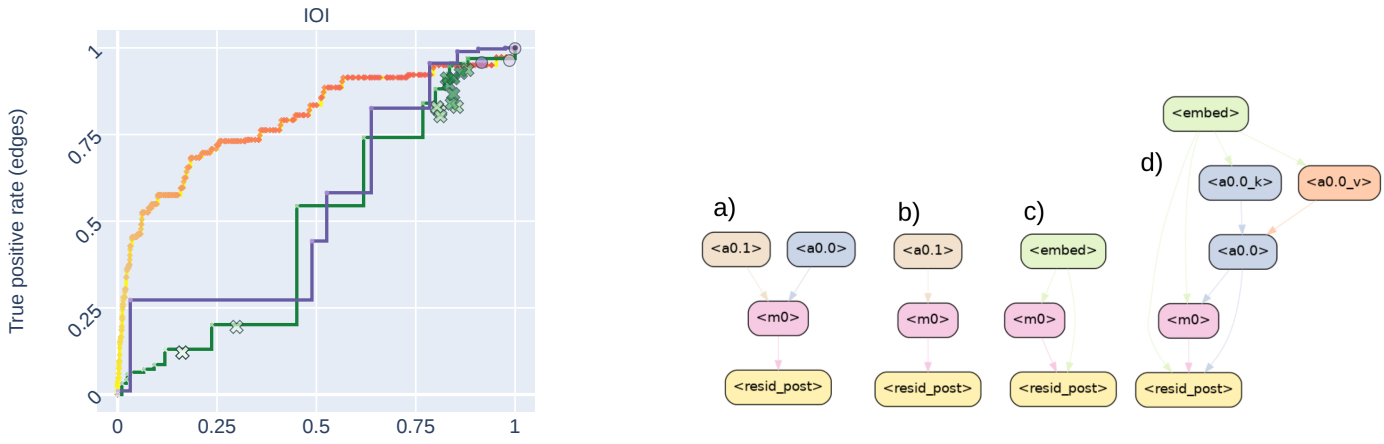


Figure 4: Comparison of ACDC and SP with both zero activations (Figure 4a) and corrupted activations (Figure 4b). We measure each subgraph's KL divergence to the whole model, on held out i.i.d test data. The darker points include more edges in the hypotheses: they use smaller ACDC τ, smaller SP regularization λ or a higher percentage of nodes in HISP.



Left: updated Figure 15 for the appendix. Right: OR gate recovery. a) the ground truth: a toy model of an OR gate, where MLP0 performs OR on the bias terms of a0.0 and a0.1. b) ACDC only recovers one OR gate inputs. c) HISP recovers neither OR gate input (and also recovers the unnecessary input node). d) SP recovers only one OR gate input, and several additional nodes.